Simple Least Squares Estimator for Treatment Effects Using Propensity Score Residuals

(2018 *Biometrika*; an example for military rank effect on wage added)

Myoung-jae Lee

Korea University

April 26, 2018

Mean Difference and PS Matching

- For a binary treatment D, a response Y and covariates X, let Y^d be the potential response for $D = d$; $Y = (1 - D)Y^0 + DY^1$. If D is randomized,

$$E(Y|D = 1) - E(Y|D = 0) = E(Y^1 - Y^0).$$

- The sample version of $E(Y|D = 1) - E(Y|D = 0)$ equals

$$\text{Slope OLS of } Y \text{ on } (1, D) = \text{OLS of } Y - \bar{Y} \text{ on } D - \bar{D}.$$

- Suppose D is not randomized and X needs to be controlled. If $(Y^0, Y^1) \amalg D|X$, then

$$E(Y|D = 1, X) - E(Y|D = 0, X) = E(Y^1 - Y^0|X).$$

- To avoid the dimension problem in controlling X, propensity score (PS) matching with $\pi(X) \equiv E(D|X)$ is used because

$$Y^d \amalg D|X \Longrightarrow Y^d \amalg D|\pi(X) \ \forall d \qquad \text{(Rosenbaum \& Rubin, 1983 BKA)}$$

Problems with Propensity Score Matching (PSM)

- PSM requires several decisions on the user, according to which the effect estimate can change much.
- First, how many matched subjects for individual i: one for pair matching, and more for multiple matching.
- Second, whether to use a fixed number of matches M, or an individual-varying number M_i.
- Third, whether to use a caliper (a bound on the deviation between X_i and X of a matched individual) or not; if yes, its value.
- Fourth, matching with replacement or without. And more,...
- Getting standard errors in PSM is hard, despite the asymptotic normality in Abadie & Imbens (2016, ECA) under a parametric $\pi(X)$.
- The variance estimator is complicated, involving

$$V(Y|D=d,\pi(X)=p) \quad \& \quad COV\{X, E(Y|D=d,X)|\pi(X)=p\}.$$

Matching is a Threat to Public Health?

- PSM has been applied at least 55 times in four best medical journals (Wu et al. 2015, Epidemiology): New England Journal of Medicine, Lancet, Journal of the American Medical Association, and British Medical Journal.
- Among them, 21 studies used caliper matching, in which 10 studies used 0.2 SD(logit PS) as the caliper, and 4 studies used 0.6 SD; 1:1 matching was the most popular, although it is inferior to multiple matching.
- Among 55 studies, only 44 (80%) checked the covariate balance after PSM, including only 21 studies that (48%) did statistical tests.
- Many people do not listen to economists, who lost much credibility in the wake of the financial crises, but they still adhere to medical advices in those leading journals in clinical medicine.
- So PSM could be a matter of life or death, and yet arbitrariness and lack of statistical prowess exist in PSM applications, which is disheartening, and worrying as well for public health.

Main Idea of PS-Residual (PSR) OLS

- Is it possible to bring back the simple OLS of Y on $(1, D)$ while still controlling X nonparametrically?
- Under $Y^d \amalg D|X$ and the support-overlap condition $0 < \pi(X) < 1$, the answer is positive: do

$$\text{OLS of} \quad Y - \bar{Y} \quad \text{on} \quad D - \hat{\pi}(X). \qquad (\text{OLS}^0_{psr})$$

- OLS^0_{psr} includes the simple OLS for randomized D as a special case, because $\pi(X) \equiv E(D|X) = E(D)$; the superscript 0 will be explained shortly.
- It may look puzzling why X does not appear as regressors along with $D - \hat{\pi}(X)$. *The key point is that X is uncorrelated with $D - \pi(X)$, and thus X can be put into the error; balancing/matching on X unnecessary.*
- If $\pi(X)$ is estimated nonparametrically, OLS^0_{psr} is nonparametric because the X-part not specified. With probit used for $\pi(X)$ under $\pi(X) = \Phi(X'\alpha)$ in this paper, OLS^0_{psr} is semiparametric.

Generalizing PSR-OLS

- Let $\Pi^q(Y|X'\alpha)$ denotes the linear projection of Y on $\{1,\ X'\alpha,\ ...,\ (X'\alpha)^q\}$. A generalized version of OLS^0_{psr} is

$$\text{OLS of} \quad Y - \Pi^q(Y|X'\alpha) \quad \text{on} \quad D - \pi(X).$$

- Denoting the projection coefficient for $(X'\alpha)^j$ as γ_j, $j = 0, ..., q$, estimate the γ_j's with the OLS of Y on $\{1,\ X'\hat{\alpha},\ ...,\ (X'\hat{\alpha})^q\}$, where $\hat{\alpha}$ is the probit of D on X (i.e., $\hat{\pi}(X) = \Phi(X'\hat{\alpha})$). Let $\gamma \equiv (\gamma_0, \gamma_1, ..., \gamma_q)'$.

- Implement the above OLS with (set q at $2, 3$ in practice)

$$\text{OLS of} \quad Y - \sum_{j=0}^{q} \hat{\gamma}_j (X'\hat{\alpha})^j \quad \text{on} \quad D - \Phi(X'\hat{\alpha}). \qquad (\text{OLS}^q_{psr})$$

- OLS^q_{psr} includes OLS^0_{psr} as a special case when $q = 0$. To ease referencing OLS^0_{psr} and OLS^q_{psr} with $q > 0$, however, use the expression OLS^q_{psr} only for $q > 0$ henceforth. 'OLS_{psr}' refers to both OLS^0_{psr} and OLS^q_{psr}.

Advantages of PSR-OLS and Remarks

- First, OLS_{psr} is possibly the easiest to implement, with hardly any choice required by the user; it is numerically stable.
- Second, it has a simple asymptotic variance estimator that works also well in small samples.
- Third, as will be seen, it can be easily extended to multiple/multi-valued D by replacing $\pi(X)$ with a 'generalized PS'.
- The motivation to extend OLS^0_{psr} to OLS^q_{psr} is to improve OLS^0_{psr} in case PS is misspecified, although OLS_{psr} proceeds on the premise of the correctly specified PS as PSM does—more on this shortly.
- Simply put, OLS_{psr} brings the "time-tested work horse" OLS back to life for binary or multiple treatment while controlling covariates semiparametrically.

Motivating Semi-Linear Parallel-Shift Model

- For an unknown $\mu(\cdot)$, let (this "parallel shift" will be relaxed later):

$$Y = \beta D + \mu(X) + U \qquad \text{where} \qquad E(U|X) = 0.$$

- Note $U \amalg D|X \Longrightarrow U \amalg D|\pi(X)$. Take $E\{\cdot|\pi(X)\}$ on the Y eq.:

$$E\{Y|\pi(X)\} = \beta\pi(X) + E\{\mu(X)|\pi(X)\};$$

take $E\{\cdot|\pi(X)\}$ on $\pi(X) \equiv E(D|X)$ to see $\pi(X) = E\{D|\pi(X)\}$.

- Subtract $E\{Y|\pi(X)\}$ eq. from Y eq., and subtract & add $E(Y)$ to get

$$Y - E(Y) = \beta\{D - \pi(X)\} + V \qquad \text{where} \quad V \text{ is '} U + \text{function of } X\text{'}$$

- OLS^0_{psr} works due to $U \amalg D|X \Longrightarrow V \amalg D|X \Longrightarrow V \amalg D|\pi(X)$ implying

$$E[\{D - \pi(X)\}V] = E[\ E\{DV - \pi(X)V|\pi(X)\}\ ] = 0.$$

Implementation and Generalization

- With $\pi(X) = \Phi(X'\alpha)$ to apply probit for α, OLS^0_{psr} is much easier to implement than PSM.
- When PS is misspecified, $COR\{D - \pi(X), V\} \neq 0$ in general, and the omitted X-dependent terms in V result in biases. This may be alleviated if $E\{Y|\pi(X)\}$ is explicitly accounted for by $\Pi^q(Y|X'\alpha)$ in OLS^q_{psr}.
- Using $X'\alpha$ instead of $\Phi(X'\alpha)$ in $\Pi^q(Y|X'\alpha)$ makes the extension to multiple treatments easier.
- In OLS_{psr}, the only decision to make is specifying the PS regression function $X'\alpha$, which is common for all PS-based estimators. For simplicity, proceed with OLS^2_{psr} henceforth, unless otherwise noted.

Asymptotic Distribution

- With $\pi(X)$ and $E(Y|X)$ nonparametrically estimated in the OLS of $Y - E(Y|X)$ on $D - \pi(X)$, the first-stage errors, $\hat{\pi}(X) - \pi(X)$ and $\hat{E}(Y|X) - E(Y|X)$, are orthogonal to the OLS moment condition.

- But for OLS^2_{psr} denoted as $\hat{\beta}^2_{psr}$, the error $\hat{\alpha} - \alpha$ matters, and

$$\sqrt{N}(\hat{\beta}^2_{psr} - \beta) \rightsquigarrow N(0, \Omega) \quad \text{where} \quad \hat{\Omega} \equiv \left(\frac{1}{N}\sum_i \hat{\varepsilon}_i^2\right)^{-2} \cdot \frac{1}{N}\sum_i (\hat{V}_i\hat{\varepsilon}_i + \hat{L}\hat{\eta}_i)^2$$

and

$$\begin{aligned}
\hat{\varepsilon}_i &\equiv D_i - \Phi(X_i'\hat{\alpha}), \qquad \hat{V}_i \equiv Y_i - \{\hat{\gamma}_0 + \hat{\gamma}_1 X_i'\hat{\alpha} + \hat{\gamma}_2 (X_i'\hat{\alpha})^2\} - \hat{\beta}^2_{psr}\hat{\varepsilon}_i, \\
\hat{\eta}_i &\equiv \left(\frac{1}{N}\sum_i \hat{s}_i\hat{s}_i'\right)^{-1}\hat{s}_i \qquad \text{with} \qquad \hat{s}_i \equiv \frac{\hat{\varepsilon}_i \phi(X_i'\hat{\alpha})}{\Phi(X_i'\hat{\alpha})\{1 - \Phi(X_i'\hat{\alpha})\}} X_i, \\
\hat{L} &\equiv -\frac{1}{N}\sum_i \hat{V}_i \phi(X_i'\hat{\alpha}) X_i'.
\end{aligned}$$

- If more polynomial terms of $X'\alpha$ are used for $\Pi^q(Y|X'\alpha)$, the modification needed is adding the extra terms into $\hat{V}$; $\hat{V}_i \equiv Y_i - \bar{Y} - \hat{\beta}^0_{psr}\hat{\varepsilon}_i$ in OLS^0_{psr}.

Efficiency Question and Remarks

- The simulation section will demonstrate that $\hat{\Omega}$ works well in small samples. If desired, use nonparametric bootstrap, resampling from the original sample with replacement.

- Hahn (1998 ECA, p. 323) showed that the OLS of $Y - E(Y|X)$ on $D - E(D|X)$ is *not* semiparametrically efficient. This suggests that, with α further estimated, OLS_{psr} would not be semiparametrically efficient.

- Despite the inefficiency, a simulation study will show that, in finite samples, OLS_{psr} is far more efficient as well as less biased than supposedly efficient estimators.

- This holds despite no user-interventions on OLS_{psr}, such as using a caliper in matching or excluding extreme observations with $\pi(X) \simeq 0, 1$ in weighting.

- An informative way to view OLS_{psr} is that it alters the treatment dose from $0, 1$ to $D - \pi(X) \in (-1, 1)$. E.g., if a person with $\pi(X) = 0.9$ is treated, the "effective dose" is only $1 - 0.9 = 0.1$ to compensate for the PS imbalance.

General Model with Heterogeneous Effect

- To relax the parallel shift, let, for unknown $\mu(X)$ & $\mu_D(X)$,

$$Y = \mu(X) + \mu_D(X)D + U \implies E(Y^1 - Y^0|X) = \mu_D(X).$$

- In this general-shift, which becomes the parallel shift under $\mu_D(X) = \beta$,

$$\hat{\beta}_{psr} \to^p \beta_\omega \equiv E\{\omega(X)\mu_D(X)\} = E\{\omega(X)E(Y^1 - Y^0|X)\}$$

where

$$\omega(X) \equiv \frac{\pi(X)\{1-\pi(X)\}}{E[\pi(X)\{1-\pi(X)\}]} = \frac{V(D|X)}{E\{V(D|X)\}}.$$

- $\beta_\omega \neq \beta \equiv E(Y^1 - Y^0) = E\{E(Y^1 - Y^0|X)\}$, but $\beta_\omega = \beta$ if parallel shift, or if the $\mu_D(X)$-covariates are independent of the $\pi(X)$-covariates due to

$$E\{\omega(X)\mu_D(X)\} = E\{\omega(X)\} \cdot E\{\mu_D(X)\} = E\{\mu_D(X)\}.$$

Why the Weighted Effect is Good

- When the X-conditional effect is $\mu_D(X)$, for the population, it is a matter of how to average X out. In a weighted averaging, higher weights are given to individuals deemed to be more important for the research or policy purpose.

- This importance is gauged by $f_X(X)$ in $E\{\mu_D(X)\}$, and by $\omega(X)f_X(X) \propto \pi(X)\{1-\pi(X)\}f_X(X)$ in $E\{\omega(X)\mu_D(X)\}$.

- Since $\{1-\pi(X)\}\pi(X)$ attains its max. at $\pi(X) = 0.5$ & decreases to 0 as $\pi(X) \to 0, 1$, those with $\pi(X) \simeq 0.5$ get higher weights in $E\{\omega(X)\mu_D(X)\}$ (& those with $\pi(X) \simeq 0, 1$, lower weights). Why is this good?

- First, those with $\pi(X) \simeq 0.5$ are close to being randomized, thus less susceptible to confounding by unobservables; they deserve high weights.

- Second, other estimators have an arbitrary feature to downweight extreme observations with $\pi(X) \simeq 0, 1$, but the $\omega(X)$-weighting of OLS_{psr} is a built-in, non-arbitrary feature to downweight observations with $\pi(X) \simeq 0, 1$.

Non-Continuous Response

- OLS_{psr} works for any response Y, not just continuously distributed Y.
- E.g., suppose $Y = 1[X'\psi + \beta D + N(0,1) > 0]$. Then,

$$\begin{aligned} \mu(X) &= \Phi(X^T\psi), \quad U^0 = Y - \Phi(X^T\psi), \\ \mu_D(X) &= \Phi(X^T\psi + \beta) - \Phi(X^T\psi), \quad U^1 = Y - \Phi(X^T\psi + \beta). \end{aligned}$$

- $\hat{\beta}_{psr} \to^p E[\omega(X)\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$, while typically $E[\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$ is presented as a marginal effect.
- For Y probit, estimating $E[\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$ requires an extra work. In contrast, $\hat{\beta}_{psr} \to^p E[\omega(X)\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$ directly, with an extra work done for the D probit instead.
- This is fine as long as misspecifications in $\pi(X)$ are less worrisome than those in the Y-model, which is the stance taken in the PSM literature, as it has chosen to specify $\pi(X)$, instead of $E(Y^d|X) = E(Y|D = d, X)$.

Weighted PSR-OLS

- Consider the OLS $\hat{\beta}^{\omega}_{psr}$ for

$$\frac{Y - E\{Y|\pi(X)\}}{[\pi(X)\{1-\pi(X)\}]^{1/2}} \quad \text{on} \quad \frac{D - \pi(X)}{[\pi(X)\{1-\pi(X)\}]^{1/2}}$$

- As $\omega(X)$ is removed by the weighting,

$$\hat{\beta}^{\omega}_{psr} \to^{p} \beta \equiv E\{E(Y^1 - Y^0|X)\} = E(Y^1 - Y^0).$$

  The asy.dist. of $\sqrt{N}(\hat{\beta}^{\omega 2}_{psr} - \beta)$ is similar to that of $\sqrt{N}(\hat{\beta}^{2}_{psr} - \beta_{\omega})$.

- Unless $\hat{\pi}(X)$ is well bounded within $(0, 1)$, however, the finite sample performance of $\hat{\beta}^{\omega}_{psr}$ would not be as good as $\hat{\beta}_{psr}$ due to $\hat{\pi}(X) \simeq 0, 1$.

- This can be overcome by using only observations with $\hat{\pi}(X)$ away from 0 and 1, which brings in arbitrariness though; $\hat{\beta}_{psr}$ would be preferred to $\hat{\beta}^{\omega}_{psr}$.

Multiple OLS for Multiple Treatment

- Suppose D takes on $0, 1, ..., J$. Let $D_d \equiv 1[D = d]$ to consider parallel-shift:

$$Y = \mu(X) + \sum_{d=1}^{J} \beta_d D_d + U \quad \text{where} \quad E(U|X) = 0.$$

- With $\pi_d(X) \equiv E(D_d|X)$ and $\pi(X) \equiv \{\pi_1(X), ..., \pi_J(X)\}'$,

$$Y - E(Y|\pi(X)) = \sum_{d=1}^{J} \beta_d \{D_d - \pi_d(X)\} + V.$$

- The analog for OLS^0_{psr} is

$$\text{OLS of } Y - \bar{Y} \quad \text{on} \quad D_d - \pi_d(X), \quad d = 1, ..., J.$$

- The analog for OLS^q_{psr} is

$$\text{OLS of } Y - \Pi^q(Y|X'\alpha) \quad \text{on} \quad D_d - \pi_d(X), \quad d = 1, ..., J$$

where $X'\alpha$ can be uni- or multi-dimensional; examples next.

Multiple Treatment Cases

- First, the treatments are *ordered* to be generated by

$$D_i = \sum_{d=1}^{J} 1[\zeta_d \leq X_i'\alpha + \varepsilon_i], \quad \zeta_1 = 0 < \zeta_2 <, \cdots, < \zeta_J.$$

  - E.g., D is schooling years. Under $\varepsilon \sim N(0,1) \amalg X$, apply ordered probit to estimate the 'single index' $X'\alpha$. Then use $\Pi^q(Y|X'\alpha)$.

- Second, the treatments are *partly ordered* as in (Ju & Lee, 2017 OBES)

$$\begin{aligned} D_{0i} &\equiv 1[0 \leq X_{0i}'\alpha_0 + \varepsilon_{0i}], \quad D_{ri} \equiv 1 + \sum_{d=1}^{J-1} 1[\zeta_d \leq X_{ri}'\alpha_r + \varepsilon_{ri}], \\ \zeta_1 &= 0 < \zeta_2 <, \cdots, < \zeta_{J-1}, \quad D_i \equiv (1 - D_{0i})D_{ri} \text{ taking on } 0, 1, 2, ..., J. \end{aligned}$$

  - E.g., $D_0 = 1$ if not joining military, and $D_r = 1, 2, ..., J$ is military rank. (D_0, D_r) depends on X through $(X_0'\alpha_0, X_r'\alpha_r)$. Use $\Pi^q(Y|X_0'\alpha_0, X_r'\alpha_r)$.

- Third, if D is *multinomial*, J linear indices appear; e.g., D represents job categories.

Other Estimators: Regression Imputation (RI) and PSM

- 2nd-order series-approximation 'regression imputation' (RI) estimator $\hat{\beta}_{ri2}$ is

$$\frac{1}{N}\sum_i\{\hat{\tau}_{10}+\hat{\tau}_{11}\hat{\pi}(X_i)+\hat{\tau}_{12}\hat{\pi}(X_i)^2\}-\frac{1}{N}\sum_i\{\hat{\tau}_{00}+\hat{\tau}_{01}\hat{\pi}(X_i)+\hat{\tau}_{02}\hat{\pi}(X_i)^2\};$$

  $(\hat{\tau}_{d0},\hat{\tau}_{d1},\hat{\tau}_{d2})$ is the OLS of Y on $\{1,\hat{\pi}(X),\hat{\pi}(X)^2\}$ on $D=d$. Using $\hat{\pi}(X)^3$ additionally gives $\hat{\beta}_{ri3}$.

- A PS pair-matching estimator for $E(Y^1-Y^0)$ is

$$\hat{\beta}_{m1}\equiv\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^1-\hat{Y}_i^0)\quad\text{with}\quad\hat{Y}_i^1\equiv D_iY_i+(1-D_i)Y_{t(i)}$$

$$\hat{Y}_i^0\equiv(1-D_i)Y_i+D_iY_{c(i)};$$

  $t(i)$ is the matched treated for control i; $c(i)$ matched control for treated i.

- If $Y_{c(i)}$ is replaced by the average of the four nearest controls and if $Y_{t(i)}$ is replaced by the average of the four nearest treated, then 'PS four-multiple-matching estimator' $\hat{\beta}_{m4}$ is obtained.

Other Estimators: Bias-Corrected PSM

- Whereas the above RI and PSM specify $\pi(X)$, not $E(Y^d|X) = E(Y|X, D = d)$, there are estimators specifying $E(Y|X, D = d) = X'\beta_d$ (and $\pi(X)$).

- A bias-corrected version of $\hat{\beta}_{m1}$ (Abadie and Imbens 2011, JBES) is

$$\hat{\beta}_{mbc} \equiv \frac{1}{N}\sum_{i=1}^{N}(\tilde{Y}_i^1 - \tilde{Y}_i^0), \ \tilde{Y}_i^1 \equiv D_i Y_i + (1 - D_i)(Y_{t(i)} + X_i'\hat{\beta}_1 - X_{t(i)}'\hat{\beta}_1),$$
$$\tilde{Y}_i^0 \equiv (1 - D_i) Y_i + D_i(Y_{c(i)} + X_i'\hat{\beta}_0 - X_{c(i)}'\hat{\beta}_0).$$

- Matching is not exact (i.e., $X_{t(i)} \neq X_i$ or $X_{c(i)} \neq X_i$) to cause a bias, and adding $X_i'\hat{\beta}_1 - X_{t(i)}'\hat{\beta}_1$ and $X_i'\hat{\beta}_0 - X_{c(i)}'\hat{\beta}_0$ avoids the bias.

- $\hat{\beta}_{mbc}$ differs from Abadie and Imbens (2011): $\hat{\beta}_{mbc}$ uses linear models for $E(Y^d|X)$ while Abadie and Imbens used nonparametric estimators, and $\hat{\pi}(X)$ is used in selecting $t(i)$ and $c(i)$ while X is used in Abadie and Imbens.

Simulation Study 1

- The basic simulation design is: with the simulation repetition 10000,

$$D = 1[0 < \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon], \qquad \varepsilon \sim N(0,1) \amalg (X_2, X_3),$$
$$(X_2, X_3) \text{ is jointly standard normal with } COR(X_2, X_3) = \sqrt{0.5} \simeq 0.71,$$

-

$$Y = \beta_d D + \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U, \qquad U \sim N(0,1) \amalg (X_2, X_3, \varepsilon),$$
$$\alpha_1 = 0,\ \alpha_2 = 1,\ \alpha_3 = \pm 1,\ \ \beta_1 = 0,\ \beta_d = \beta_2 = \beta_3 = 1,\ \ N = 400.$$

- $E(D) \simeq 0.5$. When $\alpha_3 = -1$, (X_2, X_3) averages around $(-0.2, 0.2)$ and $(0.2, -0.2)$ for the two groups, but when $\alpha_3 = 1$, much further away, around $(-0.7, -0.7)$ and $(0.7, 0.7)$; X overlaps much better in the former.

- 9 Estimators compared: $\hat{\beta}_{ri2}$, $\hat{\beta}_{ri3}$, $\hat{\beta}_{m1}$, $\hat{\beta}_{m4}$, $\hat{\beta}_{mbc}$, $\hat{\beta}^{2}_{psr}$, $\hat{\beta}^{4}_{psr}$, $\hat{\beta}^{\omega 2}_{psr}$, $\hat{\beta}^{\omega 4}_{psr}$.

Simulation Study 2

Table 1. Good X-Overlap & Poor X-Overlap	|bias|, sd, rmse	|bias|, sd, rmse
$\hat{\beta}_{ri2}$	0.00, 0.14, 0.14	0.86, 0.24, 0.89
$\hat{\beta}_{ri3}$	0.00, 0.16, 0.16	0.03, 0.26, 0.26
$\hat{\beta}_{m1}$	0.00, 0.23, 0.23	0.33, 0.33, 0.47
$\hat{\beta}_{m4}$	0.00, 0.17, 0.17	0.47, 0.23, 0.52
$\hat{\beta}_{mbc}$	0.00, 0.15, 0.15	0.00, 0.32, 0.32
$\hat{\beta}^{2}_{psr}$	*0.00, 0.12, 0.12*	*0.00, 0.15, 0.15*
$\hat{\beta}^{4}_{psr}$	*0.01, 0.12, 0.12*	*0.00, 0.15, 0.15*
$\overline{sd}$	0.12, 0.12	0.15, 0.15
$\hat{\beta}^{\omega 2}_{psr}$	0.01, 0.17. 0.17	0.03, 0.27, 0.27
$\hat{\beta}^{\omega 4}_{psr}$	0.01, 0.17, 0.17	0.01, 0.27, 0.27
$\overline{sd}^{\omega}$	0.17, 0.17	0.22, 0.21
$\overline{sd}$: avg. of asy. sd est. for $\hat{\beta}_{psr}$; similar for $\overline{sd}^{\omega}$		

Simulation Study 3

Table 2. Poor X-Overlap; $\hat{\beta}_m$ with Caliper 0.05; $\hat{\beta}^{\omega}_{psr}$ with $0.01 < \hat{\pi}(X) < 0.09$

	base design |bias|, sd, rmse	$\pi(X)$ wrong |bias|, sd, rmse	hetero.eff.linear |bias|, sd, rmse	hetero.eff.bi Y |bias|, sd, rmse
$\hat{\beta}_{ri2}$	0.85, 0.25, 0.89	0.18, 0.21, 0.27	0.25, 0.18, 0.30	0.04, 0.08, 0.09
$\hat{\beta}_{ri3}$	0.02, 0.26, 0.26	0.00, 0.30, 0.30	0.01, 0.18, 0.18	0.03, 0.10, 0.10
$\hat{\beta}_{m1}$	0.21, 0.30, 0.36	0.02, 0.19, 0.19	0.02, 0.18, 0.18	0.04, 0.10, 0.11
$\hat{\beta}_{m4}$	0.01, 0.17, 0.17	0.00, 0.15, 0.15	0.00, 0.14, 0.14	0.09, 0.07, 0.12
$\hat{\beta}_{mbc}$	0.00, 0.32, 0.32	0.01, 0.28, 0.28	0.00, 0.18, 0.18	0.00, 0.11, 0.11
$\hat{\beta}^{2}_{psr}$	*0.00, 0.15, 0.15*	*-0.01, 0.13, 0.13*	*0.26, 0.14, 0.30*	*0.10, 0.07, 0.12*
$\hat{\beta}^{4}_{psr}$	*0.00, 0.15, 0.15*	*-0.11, 0.13, 0.17*	*0.26, 0.14, 0.30*	*0.10, 0.07, 0.12*
$\overline{sd}$	*0.15, 0.15*	*0.13, 0.13*	*0.14, 0.14*	0.07, 0.07
$\hat{\beta}^{\omega 2}_{psr}$	0.00, 0.22, 0.22	0.08, 0.20, 0.22	0.09, 0.24, 0.26	0.05, 0.08, 0.09
$\hat{\beta}^{4}_{psr}$	0.01, 0.22, 0.22	0.30, 0.21, 0.37	0.09, 0.23 0.25	0.05, 0.08, 0.09
$\overline{sd}^{\omega}$	0.19, 0.19	0.19, 0.20	0.22, 0.22	0.07, 0.07

Military Rank Effects on Wage: Mean (SD) & OLS

	1356 Non-Veterans	1816 Veterans	OLS (t-value)
1974 wage (exp(Y))	15,941 (8,083)	15,374 (7,472)	
1974 schooling years	14.5 (2.42)	13.6 (1.93)	0.038 (8.39)
1957 parent wage	6,458 (6,111)	6,330 (5,513)	0.083 (6.36)
1957 # activities	1.40 (1.50)	1.38 (1.47)	0.014 (1.96)
1957 IQ	103 (16.0)	100 (14.5)	0.395 (6.25)
1957 father alive	0.952	0.951	-0.095 (-2.89)
1957 mother alive	0.975	0.977	-0.042 (-1.00)
1957 any religion	0.789	0.758	
1957 friend military	0.097	0.219	
1974 single	0.073	0.059	-0.190 (-3.00)
1974 married	0.875	0.895	0.104 (2.33)
private	.....	0.376	-0.020 (-0.84)
corporal	.....	0.349	0.009 (0.45)
sergeant	.....	0.202	0.008 (0.29)
officer	.....	0.073	0.165 (3.07)
Ju & Lee (2017 OBES) Data with $N = 3172$			
In OLS: $Y = \ln(\text{wage})$, $\ln(\text{parent wage})$, IQ/100; $R^2 = 0.131$			

Military Rank Effect on Wage: Various Estimates

Two regression functions from probit for military or not, and ordered probit for military rank. No caliper for matching. Bootstrap for matching and RI t-values. OLS uses 12 regressors while PSR only one; remarkably, they give similar results.

Military Rank Effect on Wage: $\hat{\beta}$ (tv) (no $\hat{\beta}^{\omega}_{psr}$, as $\omega(X)$ is unavailable)				
	Private	Corporal	Sergeant	Officer
OLS $\binom{\text{ranks}}{\text{only}}$	-0.073 (-3.06)	-0.059 (-2.67)	-0.039 (-1.36)	0.347 (6.52)
OLS	-0.020 (-0.84)	0.009 (0.45)	0.008 (0.29)	0.165 (3.07)
$\hat{\beta}^{1}_{psr}$	*-0.019 (-0.82)*	*0.007 (0.34)*	*0.007 (0.26)*	*0.174 (3.25)*
$\hat{\beta}^{2}_{psr}$	*-0.017 (-0.74)*	*0.009 (0.42)*	*0.009 (0.33)*	*0.171 (3.20)*
$\hat{\beta}^{3}_{psr}$	*-0.016 (-0.70)*	*0.011 (0.50)*	*0.012 (0.43)*	*0.169 (3.15)*
$\hat{\beta}_{m1}$	-0.014 (-0.56)	-0.002 (-0.08)	0.033 (1.16)	0.410 (1.48)
$\hat{\beta}_{m3}$	-0.012 (-0.47)	0.004 (0.16)	0.023 (0.79)	0.349 (1.11)
$\hat{\beta}_{m5}$	-0.007 (-0.29)	0.008 (0.34)	0.029 (0.99)	0.102 (0.37)
$\hat{\beta}_{m7}$	-0.009 (-0.37)	0.010 (0.43)	0.020 (0.69)	0.218 (0.90)
$\hat{\beta}_{r2}$	-0.005 (-0.23)	0.007 (0.27)	0.004 (0.12)	0.305 (0.66)
$\hat{\beta}_{r3}$	-0.012 (-0.53)	0.005 (0.21)	-0.008 (-0.18)	-0.314 (-0.25)

PSR-OLS versus OLS

- The main difference between OLS_{psr} and OLS (with X linearly controlled) is that OLS_{psr} specifies the PS equation, but not the Y equation, whereas the opposite is true for OLS.
- The PS equation is a reduced form, as only the predicted probability (PS), not α per se, is the goal, whereas the Y equation is a structural form where β is the primary goal.
- The similarity between OLS_{psr} and OLS in the empirical analysis might suggest that OLS is as good as OLS_{psr}. But that is not the case.
- First, the consequences of misspecifying a reduced form as in OLS_{psr} is likely less severe than those of misspecifying a structural form as in OLS.
- Second, specifying the Y equation becomes involved when we start adding interaction terms between D and elements of X, which becomes particularly complex for multiple D.
- Third, OLS_{psr} is good for discrete responses, which is not the case for OLS.

Conclusions

- PS matching is popular in finding the effect of a binary treatment D. But it requires several arbitrary decisions, and the asymptotic inference is difficult.
- 'OLS_{psr}' uses the projection residual of D on PS, and it reduces to the OLS of Y on $(1, D)$ if D is randomized. Extended to multiple treatments.
- First, do the probit of D on X to find $\hat{\alpha}$ for $\Phi(X'\alpha)$. Second, do the OLS of Y on a polynomial function of $X'\hat{\alpha}$, to get the linear projection $\Pi^q(Y|X'\hat{\alpha})$. Third, do the OLS of $Y - \Pi^q(Y|X'\hat{\alpha})$ on $D - \Phi(X'\hat{\alpha})$ for the desired effect.
- OLS_{psr} works far better than competitors; set q at $2, 3$ in practice. The asy.variance estimator is easy to compute, and works well in small samples.
- $\text{OLS}_{psr} \to^p E\{\omega(X)E(Y^1 - Y^0|X)\}$, where $\omega(X) \propto \pi(X)\{1 - \pi(X)\}$ gives higher weights to $\pi(X) \simeq 0.5$ ("randomized") and lower weights to $\pi(X) \simeq 0, 1$; weighted $\text{OLS}^{\omega}_{psr} \to^p E\{E(Y^1 - Y^0|X)\} = E(Y^1 - Y^0)$.
- OLS_{psr} (along with $\text{OLS}^{\omega}_{psr}$) is the easiest to use, and it works well in all aspects that matter in practice—*Simplicity is a virtue, not a "sin"*.

Epilogue

- "Mostly Harmless Econometrics" by Angrist and Pischke (2009, Princeton U. Press) is popular among practitioners—for a good reason.
- In 2016, John Rust published an essay "Mostly Useless Econometrics? Assessing the Causal Effect of Econometric Theory" in little known journal *Foundations and Trends in Accounting*.
- There are many messages in the paper, but the main message is "let's do useful econometrics"; otherwise, econometrics may become marginalized, alienating practitioners to become an irrelevant science.
- One example cited is partial identification, which led to empirical helplessness of "Nothing in, Nothing out".
- Imposing a little assumption can go a long way toward providing informative and useful scientific findings that matter to our daily life. Let's do simple & sensible things, instead of "nobody-but-a-few-can-understand" things.